MICROSOFT MOVIE STUDIO VANTURE PROJECT

A)Bom movie write up

1. BUSINESS UNDERSTANDING

1.1. UNDERSTANDING THE PROBLEM

The role assumed here is that of a Data Scientist hired to work for Microsoft. The task is to process data sampled from various movie dataframes. This will be achieved by answering the research question: what types of films are currently doing the best at the box office.This will be achieved by comparing and analyzing domestic gross revenue and foreign gross revenue per year.

1.2. PROBLEM STATEMENT

The Problem statement is to determine the types of films currently doing the best at the box office by looking at the relationship between domestic gross revenue and foreign gross revenue annually.

2. DATA UNDERSTANDING

2.1. DATA COLLECTION

The data is derived from box office dataset [https://www.boxofficemojo.com/] . The data was collected from Box Office Mojo platform to see how the film industry revenues are accumulated from different parts of the world for a given period within a year .

2.2. DATA DESCRIPTION

column	Description
Title	This column represents the the title give to each movie in the bom movies
studio	This column represents the production studio for the movie under study
Domestic gross	This is the total domestic revenue derived from the sale of the movie for a given period of time
Foreign gross	This is the total foreign revenue associated with the sale of the movie overseas
Year	The column represents the year in which the revenues accrued

The columns; domestic_gross and foreign_gross will be the target variables from which we will get the measures of central tendency ,standard deviation and variance to ascertain their relationships.The other columns will be features from which correlation of the data will be analyzed

2.3. EXPERIMENTAL DESIGN

1. Loading Datasets and Preparing the Data.
2. Data Cleaning to deal with Anomalies and Outliers.
3. Exploratory Data Analysis (Univariate).
4. Conclusions and Recommendation.

3. DATA PREPARATION

3.1. SELECTING DATA

We will use all columns relevant to the domestic_gross and foreign gross revenues as they are relevant to the study.

3.2. DATA CLEANING

This was done to ensure the Validity, Accuracy, Completeness, Consistency and Uniformity of the Data.

The first thing done was to look for null values which were found to be none.
Then convert the columns to numeric values to make them uniform and readable and minimize errors. The columns were then checked to see if they were of the appropriate types / dtypes. The data was also found to be consistent there being no duplicated data.
The process was completed by checking for outliers within the data frames.

4. DATA ANALYSIS

4.1. EXPLORATORY DATA ANALYSIS

4.1.1. UNIVARIATE DATA ANALYSIS

a) Numerical Data.

There were alot of outliers in domestic gross(178), foreign domestic(253). These outliers are too many to remove as they will affect the accuracy of the data analysis and the result could be inconclusive or incorrect. The outliers suggest that the data could possibly be data that does not have a normal distribution.

b) Categorical Data
The categories we are interested in are that of the year that classifies the period in which the revenue was collected . The year accounts for almost 2,002 rows of data. The category of studio Names is also of interest because it will show the rank of top studios based on the movie they are producing.

c) Summary statistics

Statistical information	Domestic gross	Foreign gross
mode	1500000.0	1200000.0
range	700099600.0	960499400.0
Standard deviation	76400042.47678912	138300072.9817867
variance	5836966490455182.0	1.9126910186767524e+16
Q1	665500.0	4000000.0

skewness	3.227680661937993	3.0667438949171353
median(Q2)	16399999.0	19600000.0
mean	45715294.34815185	75979668.67282717
Q3	55700000.0	76450000.0
kurtosis	14.063948767296893	10.674168641009764

d) Univariate Analysis Recommendation.

The data is heavily skewed to the right due to a large number of outliers. Keeping the outliers was the best option because this is not a normal distribution. Using the domestic gross and foreign gross columns bears the same statistical similarity hence using either inplace of each other would bear the same results.

e) Bivariate data analysis
   Numeric
   There is a strong positive among domestic_gross and foreign_gross.
   The linear correlation with pearson's coefficient is more than0.78.

B) budget movie writeup

BUSINESS UNDERSTANDING

1.1. UNDERSTANDING THE PROBLEM

The role assumed here is that of a Data Scientist hired to work for Microsoft. The task is to process data sampled from various movie dataframes. This will be achieved by answering the research question: what types of films are currently doing the best at the box office.This will be achieved by comparing and analyzing domestic gross revenue and foreign gross revenue per year.

1.2. PROBLEM STATEMENT

The Problem statement is to determine the types of films currently doing the best at the 'the numbers' by looking at the relationship between domestic gross revenue and worldwide gross revenue annually. This will also include giving a look into the production budget column for the movies.

DATA UNDERSTANDING

2.1. DATA COLLECTION

The data is derived from box office dataset [https://www.the-numbers.com/] . The data was collected from 'The numbers' platform to see how the film industry revenues are accumulated from different parts of the world for a given period within a year and how they are budgeted.

2.2. DATA DESCRIPTION

column	Column description
id	This represents the identity of columns
release_date	This represents the date in time when a movie was released
movie	This is a title associated with a certain movie
production_budget	This represents the amount budget allocated to the production of a certain movie
domestic_gross	This is the total revenue collected from the domestic consumption of a certain movie
worldwide_gross	This is the total revenue collected from the overseas consumption of a certain movie

The columns domestci_gross, worldwide_gross and production_budget holds our target variables from which we get our statistical summary. The other columns will be featured from which the relationship of the data will be analyzed.

3. DATA PREPARATION

3.1. SELECTING DATA

The data from domestic_gross, worldwide_gross and production_budget will be converted to integer data type in order to drop the dollar signs and commas in between the data values. Furthermore the data will be generally described to see the data types we are working with.

3.2. DATA CLEANING

This was done to ensure the Validity, Accuracy, Completeness, Consistency and Uniformity of the Data.

The first thing done was to look for null values and they were found to be none . The columns were then converted to numeric data types. The data was also found to be consistent there being no duplicated data.

.

4. DATA ANALYSIS

4.1. EXPLORATORY DATA ANALYSIS

4.1.1. UNIVARIATE DATA ANALYSIS

a) Numerical Data.

There were a lot of outliers in the domestic gross(463), worldwide gross (431)and production budget(604).removing these outliers would affect the accuracy of the data analysis, and yield incorrect and inconclusive data analysis. This shows that they could not be normally distributed.

b) Categorical Data
Here we are looking at the ‘movie” column in relation to the production_budget to see which movie had the highest budget and use that to compare the returns on investment later.

c) Summary statistics.

Statistical summary	domestic_gross	worldwide_gross	production_budget
mean	41873326.867001034	91487460.90643376	31587757.0965064
range	936662225.0	2776345279.0	424998900.0
Standard deviation	68240597.35690415	174719968.77890477	41812076.82694309
variance	4656779127627114.0	3.052706749010146e+16	1748249768582191.8

1st Quartile (Q1)	1429534.5	4125414.75	5000000.0
median(Q2)	17225945.0	27984448.5	17000000.0
3rd Quartile (Q3)	52348661.5	97645836.50	40000000.0
skewness	3.7589273318288816	4.4914494627865444	2.718373312433728
kurtosis	22.41884897792375	31.92804318445726	10.285923930107748
mode	0.0	0.0	20000000.0

d)univariate Analysis Recommendation
The presence of too many outliers brought about the domestic_gross and worldwide_gross data being too heavily skewed to the right. Keeping the outliers in our data was the best option as these are not normally distributed datasets.
In the future we can use production_budget to do analysis and the results compared.

e)Bivariate data analysis
Numeric
Strong positive correlation was noticed among domestic_gross and worldwide_gross. There is linear correlation with pearson's coefficients more than 0.78.

5. CONCLUSION

From Box office Mojo IFC is the top studio having generated a foreign gross revenue of 1.2million us dollars in the making of the film 'Bluebeard'. On the other hand Uni has been involved in production of most top movies like 'Dogtooth'.Both the domestic and foreign gross revenues are directly related meaning that with a high yield in domestic gross revenue projects that there would be high foreign gross revenue.

From the Numbers the top movie produced in the kingkong relasedon 31st december 2014.

6. RECOMMENDATION

The Microsoft team should get in touch with both IFC and UNIVERSAL and get to know how they are going about the film production business since they are remitting the highest gross revenues. The team should also find out that movies such as “Bluebeard” and “Kingkong” that yielded so much global revenue.

My recommendation would be the Microsoft team as they venture into film production to focus on both domestic and global markets but give much attention to the global market because there is much return on investment.